Y = f(x) + e

Y Variables

* Response, target, output, outcome, dependent variable

X

* Features, inputs, predictors, independent variables

F

* Systematic information, the model

E

* Noise

Supervised learning:

* Explicitly trying to predict responses based on predictors

Scatterplot

* Should be used if positive relationship between sales and advertising



Closer points are to model => higher R2

F = 9.102x – 34,671, std.error of .419

* Increase in x by one unit is assosciated with an average change in F by 9.102 unit
* 9.102 +/- 2 \* .4 = [8.3, 9.9]

Multiple linear regression:

* Multiple predictors for same quantitative response

Variance-bias tradeoff

* Bias: Error from erroneous assumptions (under fitting)
* Variance: Error from sensitivity to small fluctuations, modelling noise (overfitting)

Defaulted = $0.005737x - -10.09

* For a one dollar increase in balance, the log odds ratio on average increases by 0.0057 holding all other predictors constant

Accuracy

* How often we are correct

Misclassification

* How often we are wrong
* 1 – accuracy

Precision

* When we predict yes, how often are we correct

Recall (sensitivity, true positive rate)

* When its actually yes, how often do we predict yes

How k-fold cross validation is implemented

* Randomly divide set into k folds of approximately equal size
* Treat first fold as validation set, fit model on remaining folds, calculate error rate from predictions for validation set
* Repeat k times, using different fold as validation set
* Estimate for test error rate is average of k estimates

Best subset selection

* Guaranteed to have smallest RSS, unlike forward and backward stepwise

Adding new predictor to existing linear regression model will

* Increase R squared, either increasing or decreasing adjusted R squared depending on how well the new predictor explains the response

Right figures show contours of error and constraints functions for lasso and ridge regression

* Lasso is less flexible and hence will give IMPROVED ACCURACY when its increase in bias is less than its decrease in variance

To fit regression splines, you need knots. Methos to decide how many, and where

* Domain knowledge
* More knots in places where we feel function might vary rapidly, fewer where it seems more stable
* Cross validation

Regression trees

* True is on left
* Top = More important determinant

Two random forest models estimated on same data

M = p predictors and M = sqrt(p) predictors, selected at each split

* Sqrt(p) will have larger bias but more variance, as there is less predictors selected at each split, the tree will fit the data less.
* P, we consider full range of predictors at each split, which can lead to overfitting if there are few important preditors

Support vector classifier advantages versus maximum margin classifier

* In many cases, no separating hyperplane exists, meaning no MMC. Support vector is used to construct a plane that does not perfectly seperete two classes
* MMC is xtremely sensitive to change in single observation, leading to overfitted models

PCA applied to data set with n observations and p predictors

* P principal components estimated
* Length of n for each component score vector

K-means clustering

* Partition observations into K clusters such that total within-cluster variation, summed over all k clusters, as small as possible
  + Within-cluster variation increases as K increases

Deep learning

* Rising popularity from hardware, datasets, algorithm